DILAF : des dictionnaires africains en ligne et une méthodologie

Mathieu Mangeot, Chantal Enguehard

DILAF : des dictionnaires africains en ligne et une méthodologie

Chantal Enguehard
chantal.enguehard@univ-nantes.fr
Laboratoire d'Informatique de Nantes Atlantique
France

Mathieu Mangeot
Mathieu.Mangeot@imag.fr
Laboratoire d'Informatique de Grenoble
France

Introduction

Les langues africaines sont peu présentes sur la Toile, les ressources linguistiques les concernant, et en particulier les dictionnaires, sont rarissimes.

Comparées à l'anglais, ou au français, ces langues font l'objet de peu d'études linguistiques, aussi les dictionnaires éditoriaux sont-ils rares et difficilement accessibles. Du fait de la rareté des études linguistiques, il n'existe quasiment pas d'outils informatiques adaptés à ces langues, qu'il s'agisse de correcteurs orthographiques interactifs, de dictée automatique ou de traduction automatique. Pourtant, ces outils seraient utiles aux populations parfois importantes pas ou peu alphabétisées et locutrices de ces langues.

Il n'existe *a priori* pas d'empêchement endogène concernant ces langues et qui pourrait expliquer cette faible présence sur Internet. En revanche, bon nombre de difficultés sont d'ordre économique : les financements sont faibles et les personnes qualifiées pour mener les travaux scientifiques ou techniques sont peu nombreuses.

Le projet DILAF (Dictionnaires bilingues Langues Africaines – Français) développe une stratégie visant à pallier ces manques par la mise en œuvre d'une méthodologie qui, quand c'est possible, minimise les financements nécessaires. Il a pour objectif de mettre en ligne des dictionnaires éditoriaux afin de permettre à la fois la consultation des informations par des utilisateurs, et l'exploitation des données par des outils de Traitement Automatique des Langues (TAL). Il est donc indispensable de pouvoir récupérer, au préalable, des dictionnaires éditoriaux. Ce projet ne concerne donc pas les langues pour lesquelles aucun dictionnaire n'existe. Des recherches sont menées pour ces langues dont certaines sont en danger de disparition, comme le projet Kamusi [Benjamin et al. 2014] .

Cinq dictionnaires ont ainsi été convertis et mis en ligne : bambara-français, haoussa-français, kanouri-français, tamajaq-français et zarma-français. Ils sont affichés sur le site www.dilaf.org.

Après une présentation du projet DILAF, nous effectuerons une visite du site web puis évoquerons les grandes étapes de la méthodologie ainsi que leurs difficultés (par exemple, le maniement d'outils théoriques comme les expressions rationnelles). Nous évoquerons des stratégies favorisant le succès d'une mise en œuvre de cette méthodologie, par exemple par la mise en place de stages d'étudiants en informatique.

1 – Le projet DILAF

1.1 Genèse

Les cinq dictionnaires éditoriaux actuellement présentés sur le site web DILAF ont été élaborés avant l'initiation de ce projet et indépendamment de celui-ci. Lors des années 2000, la coordinatrice du

projet (et co-auteur de cet article) a réalisé des missions d'enseignement et de recherche au Niger et au Mali. Au cours de ces missions, elle a eu connaissance de la production de ces dictionnaires et a immédiatement veillé à en collecter les sources informatiques avant que celles-ci ne disparaissent. Ces dictionnaires ont effectivement été imprimés mais ils restent peu accessibles pour les populations africaines du fait de leur coût (pour le dictionnaire bambara) ou de leur faible distribution (dictionnaires produits par le projet SOUTEBA).

Ce projet est donc né de l'opportunité offerte~~s~~ par ces recueils et de la volonté de rendre ces dictionnaires accessibles pour la population et pour des applications de TAL~~N~~.

Le projet a pu effectivement voir le jour une fois son financement assuré.

1.2 Objectifs

Le site web DILAF a été réalisé dans le cadre du projet DILAF financé par le Fonds Francophone des Inforoutes de l'Organisation Internationale de la Francophonie.

Le projet initial présentait plusieurs objectifs globaux :
– lutter contre l'analphabétisme,
– faciliter l'accès aux dictionnaires, favoriser l'expression d'écrits en langue nationale et en français, principalement chez les jeunes,
– encourager la production de pages web bilingues et dans les langues nationales,
– faire bénéficier les langues nationales des outils de Traitement Automatique des Langues ~~Naturelles (TALN)~~.

Il s'agissait aussi de concourir à modifier l'image des langues nationales en contribuant à la conception d'outils modernes. Il est surprenant de constater que ces langues sont parfois considérées comme des sous-langues, ou des langues désuètes, par les locuteurs africains eux-mêmes qui constatent que leur langue est peu présente sur Internet et peu pratique à utiliser avec les ordinateurs. Il est vrai qu'il n'y a guère d'applications informatiques qui leur soient consacrées (absence de claviers adaptés, de correcteurs orthographiques, etc.). Or la première étape – pour que les outils informatiques existants soient accommodés à de nouvelles langues est la création de ressources lexicales partagées dans des formats électroniques standardisés.

Le projet décline donc des objectifs spécifiques cohérents avec les objectifs globaux :
– produire des dictionnaires au format XML compatibles avec le standard international Lexical Markup Framework (LMF) de balisage de dictionnaires ~~multilingues~~,
– distribuer les ressources produites sous contrat Creative Commons (libre de droit)
– générer des dictionnaires HTML accessibles en ligne,
– produire un document méthodologique pour capitaliser l'expérience acquise pendant ce projet.

Les dictionnaires au format XML sont produits en transformant le format des fichiers des dictionnaires sources recueillis.

1.3 Partenaires

Ce projet a été initié par des universitaires français (universités de Nantes et de Grenoble) menant des recherches dans le domaine du TAL. Il comprend plusieurs partenaires africains ayant des activités en lien avec les langues des dictionnaires recueillis :
– l'Institut National de Documentation de Recherche et d'Animation Pédagogiques (INDRAP) - Niger,
– la Direction Générale de l'enseignement de base du Ministère de l'Éducation Nationale (MEN) - Niger
– le Centre National de Ressources de l'Éducation Non Formelle (CNR-ENF) - Mali,
– le Centre National de la Recherche Scientifique et Technique (CNRST) - Burkina Faso.

Des linguistes et concepteurs de manuels scolaires[1] issus de ces institutions ont directement travaillé sur les dictionnaires recueillis : Mahamou Raji Adamou (INDRAP) a travaillé sur le dictionnaire haoussa, Maï Moussa Maï (INDRAP) sur le dictionnaire kanouri, Rakiatou Rabé (INDRAP) sur le dictionnaire zarma, Issouf Modi (MEN) sur le dictionnaire tamajaq, Soumana Kané (CNR-ENF) et Mamadou Lamine (CNRST) –sur le dictionnaire bambara (le premier élaborant des textes de présentation du dictionnaire tandis que le ~~premier~~second effectuait sa transformation).

La mise en œuvre du projet a permis de s'appuyer sur cet ensemble de savoirs. Les informaticiens ont réalisé le site web et ont formé leurs collègues linguistes au maniement d'outils et de concepts théoriques de l'informatique (comme le codage Unicode des caractères, ou encore les expressions rationnelles) tandis que les linguistes ont guidé le choix des noms de balises et ont effectué les transformations des dictionnaires avec l'aide des informaticiens.

Le site web ainsi que l'affiche qui en fait la promotion ont été conçus en faisant participer tous les partenaires lors de réunions organisées pendant des ateliers de travail ; en dehors des ateliers, des consultations fréquentes ont eu lieu en utilisant le courrier électronique. L'implication des partenaires africains a été particulièrement essentielle pour ces tâches car le site web a pour objectif d'être consulté par des locuteurs des langues des dictionnaires. Il est donc fondamental d'impliquer fortement les personnes baignant dans les cultures associées à ces langues car ils sont en mesure de signaler les sensibilités les plus saillantes et de guider les choix.

1.4 Déroulement

Ce projet a été élaboré en 2008, son financement[2] a été obtenu fin 2009. Il a débuté en 2010.

Initialement le projet prévoyait trois ateliers se déroulant à Niamey[3]. L'atelier d'initiation s'est tenu en décembre 2010. D'une durée de deux semaines, il a permis d'accroître les compétences et connaissances des linguistes en ce qui concerne l'informatique, de démarrer effectivement les conversions des dictionnaires et de commencer à concevoir le site web et sa promotion. Le second atelier devait se dérouler 5 mois plus tard et un atelier final était planifié l'année suivante.

Les aléas politiques en ont décidé autrement. Dans la soirée du 7 janvier 2011 deux jeunes français[4] ont été enlevés alors qu'il étaient dans un maquis du centre de Niamey puis tués quelques heures plus tard. Ces événements ont amené à remettre en question la tenue des ateliers à Niamey, les autorités universitaires estimant que la sécurité n'était plus garantie à Niamey.

Le projet a donc dû être réorganisé, ce qui a entraîné des décalages dans le planning initial. L'atelier intermédiaire s'est tenu en juillet 2012 à Ouagadougou (comme le nombre de personnes à déplacer était plus important, le coût de cet atelier a augmenté). l'atelier final a finalement été annulé car le budget restant était insuffisant.

2. Site web DILAF

Dans son état initial, le site web DILAF héberge cinq dictionnaires bilingues :
– dictionnaire bambara-français, Charles Bailleul, édition 1996 (11071 entrées)
– dictionnaire haoussa-français destiné à l'enseignement du cycle de base 1, 2008, Soutéba (7921 entrées)
– dictionnaire kanouri-français destiné pour le cycle de base 1, 2004, Soutéba (6914 entrées) [Ari et

1 Dans la suite de ce texte, nous évoquerons ces partenaires par le terme générique de “linguistes” (par opposition aux “informaticiens”) afin d'alléger le discours, même si ce terme ne désigne pas ~~é~~exactement les professions et compétences de chacun d'eux.

2 120 000 euros dont 68 ~~0~~000 du Fonds Francophone des Inforoutes de l'Organisation Internationale de la Francophonie. Les autres financements ont été apportés par les partenaires du projet.

3 Un grand nombre de participants au projet habitant à Niamey, organiser les ateliers en ce lieu permettait de minimiser le nombre de personnes déplacées, et donc le coût du projet.

4 Vincent Delory et Antoine de Léocour-.

al. 2012]

– dictionnaire soŋay zarma-français destiné ~~pour~~à l'enseignement du ~~l~~c cycle de base 1, 2007, Soutéba (7921 entrées)

– dictionnaire tamajaq-français destiné à l'enseignement du cycle de base 1, 2007, Soutéba (5203 entrées) [Enguehard et al. 2009]

Ces cinq dictionnaires sont largement décrits dans [Enguehard et al. 2011].

Conformément aux objectifs du projet, ces dictionnaires sont accessibles dans un premier format destiné à la consultation directe par des locuteurs des langues et dans un second format adapté à des traitements informatiques et donc à de futures applications de traitement automatique des langues. Des informations complémentaires documentant les dictionnaires sont également présentées.

Au cours de la présentations du site web nous évoquerons les arguments ayant guidé les choix effectués.

2.1 - Page d'accueil

La page d'accueil est volontairement sobre. Il mentionne de manière très visible son caractère gratuit. Cette qualité est en effet cruciale pour les internautes africains.

Les cinq boutons situés en haut de la page restent présents quelle que soit la navigation dans le site. Quatre de ces boutons permettent d'accéder à des informations annexes (voir partie XXX), le cinquième permettant de revenir à cette page d'accueil. Ainsi, le visiteur est-il toujours en mesure d'accéder facilement à l'entièreté du site.

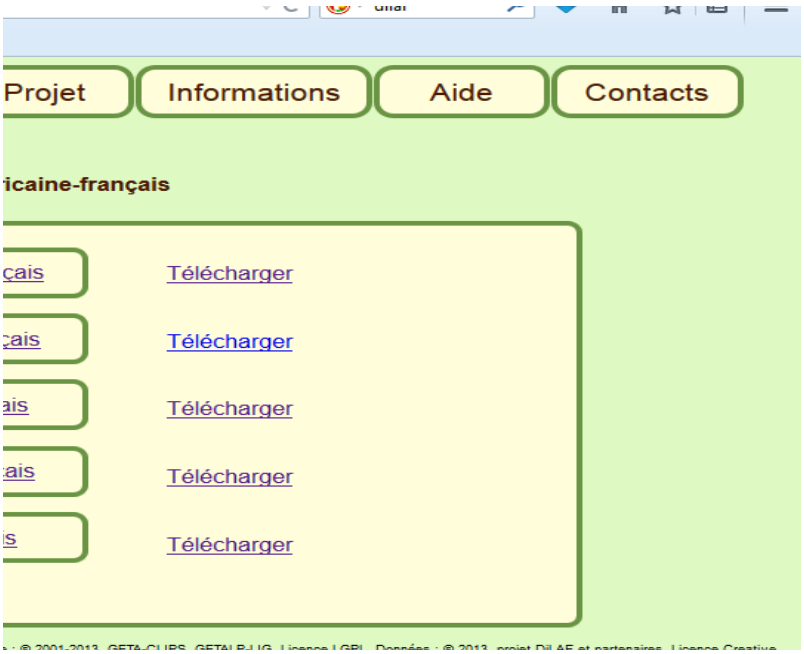

Figure 1 : Page d'accueil du site web DILAF

Cliqu~~ant~~er sur le nom des langues traitées dans les dictionnaires permet de consulter à leur contenu (voir partie 2.3). Une version XML utilisable par des applications de TAL~~N~~ peut être téléchargée (voir partie 2.5).

2.2 - Informations annexes

Le bouton "Projet" donne accès à une présentation du projet DILAF : objectifs, source de financement, ensemble des partenaires. La méthodologie DILAF est téléchargeable. Les publications scientifiques portant sur le projet sont mentionnées.

Actionner sur le bouton "Information" fait apparaître un menu déroulant mentionnant chacun des cinq dictionnaires. Cliquer sur l'un d'entre eux donne accès à un ou plusieurs fichiers. Il s'agit de textes concernant la version éditoriale du dictionnaire. Ils sont accompagnés de trois fichiers qui concernent la version en ligne du dictionnaire :-

– l'alphabet de la langue avec son ordre lexicographique et les codes Unicode de ses caractères,

– la liste des étiquettes lexicales (abréviation, forme longue et équivalent français)

– la liste des noms des éléments XML utilisés pour baliser le dictionnaire avec leur équivalent français.

Le bouton "Aide" donne quelques conseils pour utiliser la consultation en ligne.

En fin, le bouton "Contacts" indique une adresse de courrier électronique permettant de contacter les auteurs du site web.

2.3 – Consultation d'un dictionnaire

Choisir la consultation d'un dictionnaire dans la page d'accueil permet d'accéder à la page de consultation (voir figure2). Celle-ci présente alors le nom du dictionnaire ; la première entrée du dictionnaire apparaît dans le grand cadre.

Le cadre de gauche présente les entrées du dictionnaire dans l'ordre lexicographique. Cette fenêtre est destinée à inciter l'utilisateur à consulter des mots qu'il ne connaît pas en feuilletant le dictionnaire.

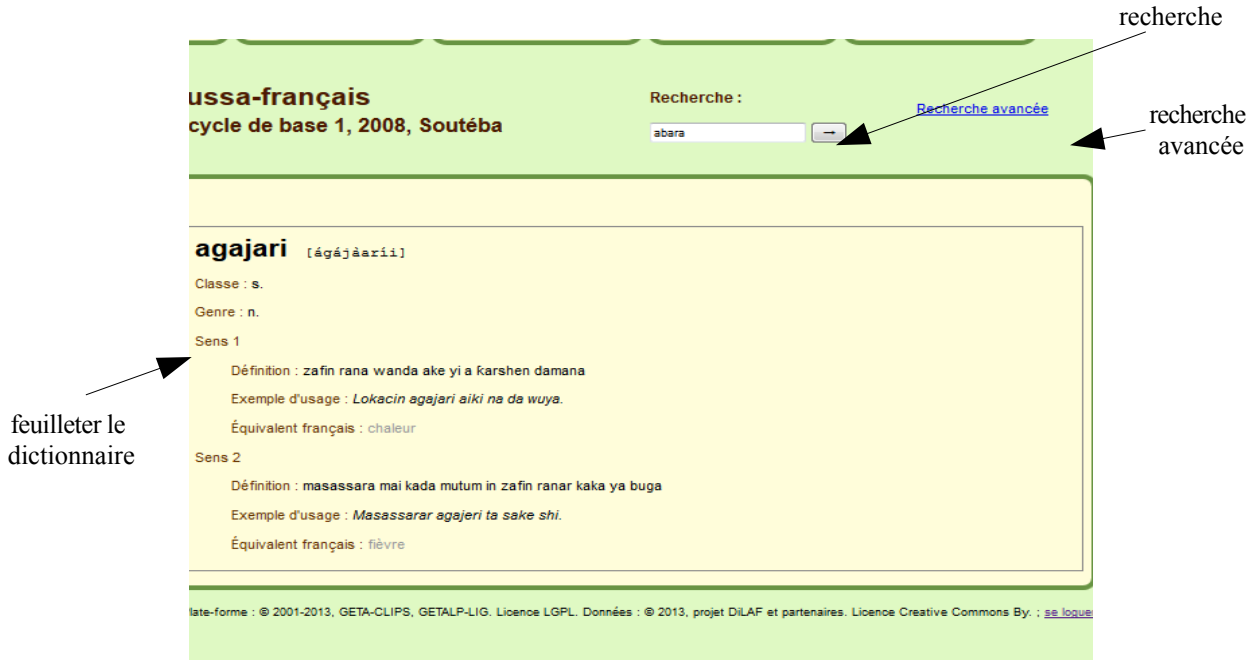

Figure 2 : Consultation de l'entrée haoussa "*agajari*"

Il est facile de rechercher un mot en en saisissant quelques lettre initiales. Un menu déroulant propose les entrées commençant par ces lettres. Si le mot saisi est trouvé dans le dictionnaire il est

affiché (s'il y a plusieurs homographes, ils sont tous affichés), et le cadre de gauche fait figurer le mot en son centre (voir figure 3). En revanche si les lettres saisies ne correspondent à aucun mot, la fenêtre est vide, mais le cadre de gauche fait apparaître les mots commençant par les lettres saisies en son centre (voir figure 4)

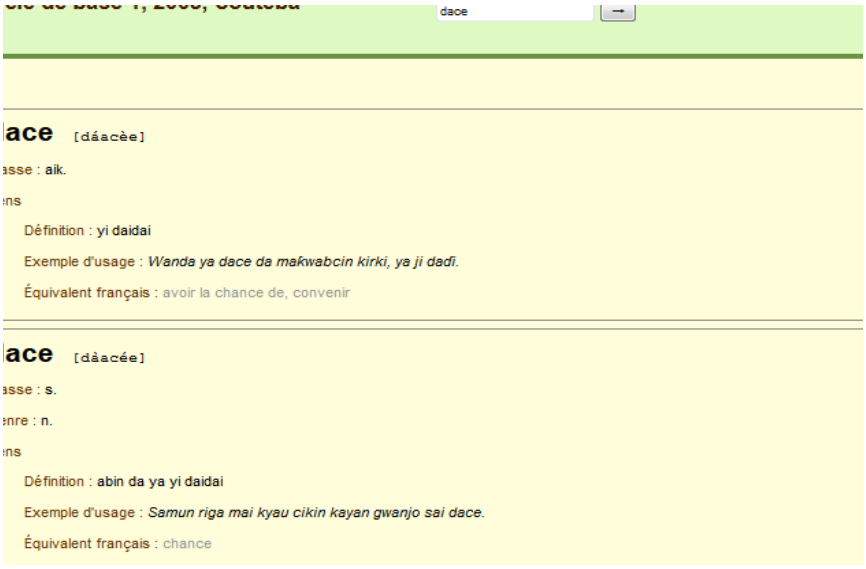

Figure 3 : La recherche de l'entrée "*dace*" fait apparaître les deux entrées "*dace*" et la fenêtre de gauche est centrée sur "*dace*"

2.4 – Recherche avancée

Une fonctionnalité de recherche avancée permet de sélectionner les entrées selon d'autres critères comme (pour le haoussa), la classe, l'équivalent français, etc. La recherche peut également porter sur plusieurs dictionnaires

2.5 – Téléchargement

Les dictionnaires au format XML peuvent être téléchargés. Ces dictionnaires sont conçus pour être utilisés par des programmes informatiques. Ils ne sont donc pas adaptés pour être lus directement par des humaines.

Le format de ces dictionnaires est largement décrit par [Enguehard et al. 2013].

Ces dictionnaires sont publiés sous licence Creative Commons "Attribution-NonCommercial-ShareAlike 2.0 Generic"[5] : les personnes ayant téléchargé un dictionnaire peuvent le modifier, l'adapter à leur guise, mais ils doivent citer l'œuvre originale ; les utilisations commerciales ne sont pas autorisées ; les œuvres modifiées peuvent être diffusées dans les mêmes conditions (avec la même licence) que l'œuvre originale.

3. Méthodologie

La méthodologie DILAF a été élaborée pour encourager la conversion de nouveaux dictionnaires éditoriaux vers des formats électroniques.

Six étapes principales ont été identifiées :

1 – Prétraitements : il s'agit de constituer un fichier OpenOffice contenant l'entièreté du dictionnaire

5 http://creativecommons.org/licenses/by-nc-sa/2.0/legalcode

(sans mise en forme des pages)

2 – Conversion d'un fichier vers Unicode : si nécessaire, les caractères spéciaux doivent être identifiés et remplacés [Enguehard 2009].

3 – Choix des noms des éléments XML : il faut nommer les informations composant un article, comme le mot-vedette, sa phonétique, la définition ou encore les synonymes. Les linguistes sont encouragés à choisir des noms dans leur propre langue afin d'utiliser cette terminologie. Si celle-ci n'existe pas, la création terminologique est encouragée.

4 – Transformations à l'aide d'expressions rationnelles : l'application d'expressions rationnelles transforme le dictionnaire de version en version jusqu'à aboutir à un dictionnaire XML entièrement balisé.

5 – Contrôle des étiquettes lexicales : dans un dictionnaire éditorial, il arrive que la même étiquette lexicale apparaisse avec de légères variations, il s'agit de rétablir un libellé unique pour chaque étiquette lexicale.

6 – Traitement des liens : Les dictionnaires éditoriaux font fréquemment apparaître des liens de synonymie, d'antonymie, etc. Il s'agit de vérifier si ces liens désignent des entrées du dictionnaire et, dans ce cas, de les faire explicitement apparaître afin ~~afin~~ qu'un internaute puisse les parcourir par des clics de souris.

Deux principes accompagnent ces étapes. Ils visent à installer de bonnes pratiques :

– Versionner : il faut nommer les différentes versions du dictionnaire au fur et à mesure de ses transformations et documenter ces transformations.

– Détecter une perte antérieure de données : il faut effectuer des contrôles aléatoires afin de détecter d'éventuelles pertes d'informations et, dans ce cas, revenir à une version antérieure.

Cette méthodologie a été testée lors du déroulement du projet DILAF, ce qui a permis de faire émerger des questionnements et de détecter les difficultés. Il est apparu que certaines étapes demandent des connaissances techniques en informatique que des personnes non professionnelles de ce domaine peuvent avoir des difficultés à maîtriser (c'est le cas des transformations du dictionnaire par des expressions rationnelles). D'autre part, certaines tâches nécessitent des compétences linguistiques professionnelles et ne pourraient être réalisées par des informati cie~~o~~ns (nommer les éléments XML par exemple).

C'est pourquoi nous conseillons de constituer une équipe avec au moins un linguiste et un informaticien pour réaliser ce travail, chacun apportant son expertise. Comme les dictionnaires éditoriaux sont réalisés par des lexicographes, il est probable que leurs fichiers soient récupérés par des linguistes. Ceux-ci pourraient initier des collaborations avec des départements d'informatique d'écoles ou d'universités proches en proposant des stages pour les étudiants.

Conclusion

Les objectifs annoncés par le projet sont atteints : les cinq dictionnaires éditoriaux sont en ligne, consultables directement par les internaute~~s~~, et également sous une forme adaptée aux traitements automatiques des langues. La méthodologie DILAF ~~–~~a été rédigée et est également librement accessible.

Bien que le projet soit terminé, beaucoup reste pourtant à faire.

Il existe des erreurs dans les dictionnaires comme des liens qui n'aboutissent pas à une autre entrée. Il

serait nécessaire de passer en revue ces liens afin de les corriger ou de créer les entrées manquantes. De plus, plusieurs dictionnaires étant des premières éditions, de nombreuses entrées pourraient être révisées.

Les dictionnaires pourraient, avec profit, être enrichis. En effet, lorsqu'il s'agit de rédiger un dictionnaire éditorial, il faut se préoccuper de l'espace nécessaire à son impression (nombre de pages). Les auteurs limitent donc la longueur de chaque article afin de ne pas aboutir à un ouvrage impubliable. Or, cette contrainte disparaît avec la publication sur support électronique. Ainsi, certaines descriptions morphologiques devraient être précisées et les exemples pourraient être traduits dans d'autres langues. C'est déjà le cas pour le dictionnaire bambara dont tous les exemples sont traduits en français (l'auteur est lui-même français). Ce corpus d'exemples constitue une ressource importante pour initier des travaux de traduction automatique.

Enfin, l'interface du site web DILAF pourrait être localisée en plusieurs langues (notamment dans les langues des dictionnaires) [Osborn 2011].

Ces futurs développements restent en suspens dans l'attente des financements nécessaires.

Références bibliographiques

[Ari et al. 2012] Ari, Abdoulkarim Chérif. Boukar, Arimi. Jarrett, Kevin Anthony. Maï, Maï Moussa. Djibir, Manoua. Koré, Taweye Aïchéta Chégou. Élaboration d'un dictionnaire bilingue kanouri-français - Kalmaram tələmyindia kanori-faransa. JEP-TALN-RECITAL 2012, Atelier TALAf 2012 : Traitement Automatique des Langues Africaines, pages 13–26, Grenoble, 4 au 8 juin 2012.

[Benjamin et al. 2014] Benjamin, Martin. Radetzky, Paula. Small Languages, Big Data: Multilingual Computational Tools and Techniques for the Lexicography of Endangered Language. 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden, 26-30 April 2014.

[Enguehard 2009] Enguehard, C. Les langues d'Afrique de l'Ouest : de l'imprimante au traitement automatique des langues, Sciences et Techniques du Langage, 6, p.29-50, 2009. (ISSN 0850-3923)

[Enguehard et al. 2009] Enguehard, C., Modi I. Towards an electronic dictionary of Tamajaq language in Niger. 12th Conference of the European Chapter of the Association for Computational Linguistics EACL-09. W07 Workshop Language Technologies for African Languages. Athens, Greece, 31 March 2009.

[Enguehard et al. 2011] Enguehard, C., Kané, S., Mangeot, M., Modi I., Sanogo M.L. Vers l'informatisation de quelques langues d'Afrique de l'Ouest. 4ème atelier international sur l'Amazighe et les Nouvelles Technologies, 24 et 25 février 2011, IRCAM, Rabat, Maroc.

[Enguehard et al. 2013] Enguehard, C., Mangeot, M. LMF for a selection of African Languages. Chapter 7, book "LMF: Lexical Markup Framework, theory and practice", Ed. Gil Francopoulo, Hermès science, Paris, France, 17 p., 2013.

[Mangeot et al. 2011] Mangeot M., Enguehard, C. Informatisation de dictionnaires langues africaines-français. Actes des journées LTT 2011, Villetaneuse, 15-16 septembre 2011.

[Osborn 2011] Osborn, Don. Les langues africaines à l'ère du numérique : défis et opportunités. Presses de l'Université Laval (PUL), Canada, 2011. ISBN-10: 2763791611. ISBN-13: 978-2763791616

Financement

Le projet DiLAF est financé par le Fonds Francophone des Inforoutes de l'Organisation Internationale de la Francophonie[6].

6 http://www.francophonie.org/